

**CHAPITRE VI**  
**PHILOSOPHIE DE LA PENSÉE**  
**LOGICO-MATHÉMATIQUE**  
*Logique, raisonnement et normes de rationalité*

*"D'où la logique est-elle née dans la tête des hommes ?  
 Certainement de l'illogisme dont le domaine a dû être immense à l'origine."  
 Nietzsche*

## 1. LA NORMATIVITE DE LA LOGIQUE

On prête en général aux vérités et aux inférences logiques la propriété d'être nécessaires : si une proposition logique est vraie, alors elle *doit* être vraie, et si une inférence logique est valide alors la conclusion *doit* s'ensuivre des prémisses. Les philosophes ont cherché à expliquer ce *doit* logique de diverses manières. Certains ont soutenu que les vérités et les inférences logiques sont nécessaires parce qu'elles décrivent un univers de faits "platoniciens" indépendants de notre connaissance, d'autres ont soutenu qu'elles sont *a priori*, ou vraies indépendamment de notre expérience, parce qu'elles sont des vérités de raison immédiatement évidentes ou encore parce qu'elles sont *analytiques* ou vraies en vertu de la signification des termes qui y figurent. D'autres philosophes ont nié que les vérités logiques soient nécessaires au sens où l'on ne pourrait jamais les réviser, et ont refusé de les distinguer nettement des vérités empiriques. D'autres encore ont soutenu qu'elles étaient seulement le produit de conventions que nous adoptons, et que nous serions libres de modifier. Ces questions sont au centre de la philosophie et de l'épistémologie de la logique (Engel 1989). Mais quelle que soit la manière dont on y répond, il faut encore expliquer pourquoi le *doit* logique apparaît si contraignant, si "dur" (Wittgenstein 1956), c'est-à-dire pourquoi, si une proposition est vraie logiquement, nous *devons* la reconnaître comme telle, et pourquoi, si une inférence est valide, nous *devons* inférer la conclusion des prémisses. Quelle que soit la conception que l'on ait de la place des vérités logiques au sein de la connaissance, il faut expliquer en quoi les vérités logiques s'imposent à nous comme évidemment *correctes*, en quoi elles sont

*normatives\**. Nous pouvons formuler cette question ainsi : pourquoi *croyons-nous* aux vérités logiques, et pourquoi suivons-nous les règles de la logique ? On peut certes répondre que c'est parce que nous les avons apprises, et que les humains ne sont pas spontanément logiciens hors des cours de logique et indépendamment de la lecture des manuels de logique, comme le montre le fait qu'ils soient fréquemment enclins à faire des paralogismes. C'est précisément parce que de tels mésusages de la raison et de la logique étaient possibles que le fondateur de cette discipline, Aristote, a éprouvé le besoin de la créer et d'ajouter à son *Organon* un volume spécialement consacré aux paralogismes, avec ses *Réfutations sophistiques*. La réponse usuelle du logicien à la question qui précède est que si nous ne suivons pas les règles de la logique, nous *devrions* néanmoins les suivre. Mais pourquoi le devrions-nous ? Parce que, dit-on encore, ce sont des règles de rationalité, qu'aucun individu rationnel ne peut ni ne doit ignorer. Mais en ce cas, à nouveau, pourquoi y a-t-il une telle différence entre ce que nous croyons en fait (nos comportements logiques effectifs) et ce que, en droit, nous devrions croire (les normes de la rationalité\* logique) ? Quelle est la relation entre la psychologie du raisonnement logique et les canons de ce raisonnement, tels que les établissent les logiciens ? Les travaux contemporains de psychologie du raisonnement\* ont permis de renouveler considérablement ce vieux problème, en montrant d'une part quelle était l'étendue des erreurs systématiques que paraissent faire les sujets dans leur pratique courante du raisonnement, et en proposant diverses explications de ces déviations par rapport aux modèles normatifs\* de rationalité\* proposés par les logiciens. Cela conduit à s'interroger à nouveaux frais sur la valeur de ces modèles et sur leur normativité\* supposée.

## 2. PSYCHOLOGISME ET ANTI-PSYCHOLOGISME

La réponse la plus radicale à la question de la normativité\* de la logique consiste à séparer totalement le *doit* du *est* logique, c'est-à-dire à soutenir qu'il ne peut y avoir, par principe, de relation entre ce que la logique nous prescrit de croire et ce que nous croyons en fait, entre les lois et les règles de la logique et le comportement psychologique déductif des

agents. Soutenir qu'il puisse exister une telle relation, selon cette réponse radicale, c'est admettre l'existence d'un lien entre la vérité et la reconnaissance de la vérité. Mais les lois logiques sont vraies que nous les reconnaissons ou non comme telles. Par conséquent, selon la conception hyper-réaliste de la vérité logique, il n'est pas possible d'inférer quoi que ce soit sur la nature des lois logiques à partir de la psychologie du raisonnement des agents humains. La version la plus influente de cet argument fut fournie par Frege et Husserl au tournant de ce siècle, dans leur fameuse critique du "psychologisme"\* en logique ( Frege 1983, Husserl 1901). Sous ce nom, Frege et Husserl attaquaient les philosophes et psychologues du XIXème siècle qui prétendaient réduire les lois de la logique à des lois de la psychologie humaine, en vertu d'une psychologie associationniste selon laquelle les lois de la pensée abstraite peuvent être dérivées de lois empiriques et de relations telles que l'association des idées par contiguité, relation et généralisation (cf. Mc Namara 1986). Or une telle réduction, selon Frege et Husserl, revient à détruire le caractère de nécessité et d'objectivité des lois logiques, qui deviennent ainsi contingentes, variables, et relatives à la psychologie subjective des individus. Selon la conception platonicienne et hyper-réaliste des lois logiques défendue par Frege, les lois logiques sont au contraire des "lois de l'être-vrai" qui valent indépendamment de la manière dont nous pouvons les reconnaître comme telles. Selon Husserl, elles ont un caractère d'"idéauté" que doit présupposer quiconque, y compris le psychologue, cherche à rendre compte de leur caractère rationnel, c'est à dire de leur validité et de leur acceptabilité universelle. L'argument anti-psychologiste\* est correct : les lois logiques ne sont pas des lois variables ni subjectives, et on doit toujours distinguer le fait de *reconnaître* qu'une proposition est vraie ou valide du fait que cette proposition soit en elle-même vraie ou valide. S'il n'était pas possible de faire cette distinction, on ne pourrait pas comprendre en quoi la logique est normative\*, et peut nous imposer, dans certains cas, de croire et d'inférer ce que nous ne sommes pas, spontanément, tentés de croire et d'inférer. Mais l'argument anti-psychologiste\* est trop fort, ou prouve trop : du fait qu'il faille distinguer la vérité de la reconnaissance de la vérité, il ne s'ensuit pas qu'il n'y ait aucun lien entre la vérité logique et la psychologie des agents, ou que ce lien soit inexplicable. Dans une conception réaliste platonicienne comme

celle de Frege, le lien entre les vérités logiques et la connaissance que nous en avons est parfaitement mystérieux : il est le produit d'une "intuition" ou d'un acte cognitif de "saisie" dont on ne voit pas, étant donné le caractère transcendant de ces vérités, comment un humain pourrait l'avoir. Même si nous admettons que la logique est objective en un sens dont le psychologisme\* ne peut pas rendre compte, il nous faut encore comprendre en quoi elle peut l'être *pour nous* et non pas seulement en soi. La position platonicienne repose en fait sur une confusion quant à la normativité\* des lois logiques : elle assimile cette normativité à la vérité (absolue) des propositions correspondantes. Mais ce n'est pas (seulement) parce qu'une proposition est vraie que nous devons la croire, ou qu'elle acquiert pour nous le statut d'une norme\* : c'est (aussi) parce que nous avons décidé de lui conférer ce statut et de la traiter comme une règle\*(Bouveresse 1987). Cela dépend certes du type de vérité objective qu'exprime la proposition (n'importe quelle proposition ne peut pas devenir une norme), mais aussi de l'état de notre psychologie et de nos décisions.

La psychologie du raisonnement du vingtième siècle, de Bühler à Piaget et au "cognitivisme", a à la fois admis et contourné l'argument fregeo-husserlien: d'un côté les psychologues ont renoncé à l'idée de *réduire* les lois logiques à des "lois de la pensée" naturelles et variables, et de l'autre ils n'ont pas cessé d'analyser les processus, les schèmes et les modes de représentation que les humains utilisent quand ils font des inférences, qu'elles soient ou non sanctionnées par la logique. Ces investigations n'impliquent en rien l'adoption, par les psychologues du raisonnement, d'une thèse psychologiste\* au sens où l'entendaient Frege et Husserl, puisqu'il ne s'agit pas pour eux de nier l'objectivité et l'"idéalité" des lois logiques, ni le caractère normatif\* des théories et des systèmes des logiciens, mais plutôt, à partir de ces théories constituées, d'en envisager la "réalité psychologique", c'est-à-dire de savoir si les structures établies par la logique peuvent aussi correspondre à des structures de représentations et de processus mentaux dans l'esprit. Ce projet n'est pas incompatible avec celui de fournir une analyse de la genèse psychologique de la pensée logique, et par là de ses formes dans la logique objectivée, comme c'est le cas dans la psychologie génétique piagétienne et dans certains travaux d'inspiration évolutionniste\* contemporains (cf. *infra*). Mais il n'implique

pas, par lui-même, le type de réductionnisme que visaient les psychologues du XIX<sup>ème</sup> siècle, pas plus qu'il n'implique la confusion dénoncée par Frege et Husserl entre l'origine psychologique des vérités logiques et la nature de ces vérités même.

L'étude empirique de la réalité psychologique des schèmes d'inférence conduit cependant à un paradoxe. Elle révèle, d'une part, que les adultes humains possèdent, à un degré élevé, la capacité de raisonner déductivement et inductivement. L'espèce humaine est hautement intelligente, possède un cerveau capable de traiter rapidement un nombre considérable d'informations et a développé un système puissant de langage naturel qui a lui-même fourni des pouvoirs étendus de représentation cognitive et de communication. Mais cette étude révèle aussi, d'autre part, que les individus se trouvent très souvent incapables d'accomplir un certain nombre de tâches de raisonnement très simples, et font des erreurs apparemment systématiques dans la réalisation de ces tâches. Ce paradoxe soulève deux problèmes étroitement liés. Le premier, que l'on peut considérer comme interne à la modélisation psychologique des processus cognitifs du raisonnement logique, est celui de savoir comment les diverses théories psychologiques peuvent rendre compte des erreurs de raisonnement : comment, en particulier, les sujets peuvent-ils se tromper dans leurs performances inférentielles si l'on suppose qu'ils ont acquis et intériorisé des schèmes de raisonnement *valides* ? Le second problème est plus général et philosophique : dans quelle mesure les erreurs systématiques de raisonnement remettent-elles en cause la rationalité\* des agents humains ? Comment la logique peut-elle être une théorie de la rationalité\* normative\* si l'on établit expérimentalement que les agents ne suivent pas réellement les règles de la logique, ou même les violent régulièrement ? Si l'on devait répondre par la négative à cette question, on aboutirait à un divorce entre le *doit* et le *est* logique bien plus radical encore que celui que les anti-psychologues\* prétendaient mettre en valeur.

### 3. LE PROBLEME DE LA RATIONALITÉ INFÉRENTIELLE

Quand on pose la question : "Les humains sont-ils rationnels dans leurs raisonnements\* ?", il importe d'abord de bien distinguer divers sens de la notion, souvent vague, de rationalité. Il y a au moins deux sens importants de cette notion dans le contexte des études sur le raisonnement\*, pour chacun desquels il existe une théorie normative\* correspondante (cf Evans 1992, mais cf. également Elster 1982, qui distingue 16 sens différents du terme "rationalité"\*!). Un premier sens du terme "rationnel"\* est celui de la rationalité\* *instrumentale* des moyens en vue d'obtenir une certaine fin : on dit en ce sens qu'un agent est rationnel\* s'il "maximise son utilité espérée" dans ses choix et dans ses actions, c'est-à-dire agit toujours de manière à faire ce qui est, à ses yeux, le plus utile, en fonction de ce qui lui semble le plus probable. La théorie normative\* correspondante est la théorie de l'utilité et de la décision, qui s'appuie sur la notion ("bayésienne") de probabilité subjective d'un agent (cf. par exemple Jeffrey 1965). Ce premier sens peut également être associé à l'idée que l'espèce humaine, comme les autres espèces, s'adapte, en vertu du processus évolutif\*, à son environnement, et que la sélection naturelle "choisit", en ce sens, les dispositifs cognitifs optimaux (cf. ci-dessous § 4). En un second sens, la rationalité est la propriété qu'on attribue à un ensemble de croyances d'être cohérentes ou non contradictoires et d'être closes sous une relation de déduction. C'est le sens proprement *logique* de la rationalité\*, d'après lequel un système cognitif est rationnel\* si l'on peut lui attribuer la capacité de faire des inférences et des raisonnements logiques déductifs. La théorie normative\* correspondante est ici la logique déductive classique, sous ses variantes traditionnelles (syllogistique et calcul des propositions) ou modernes (théorie de la quantification du premier ordre). Un autre type de capacité logique est la capacité à faire des inférences inductives, qui est à la base de l'apprentissage. Dans la mesure où la rationalité instrumentale fait appel à la notion de probabilité ou de degré de croyance subjective, on peut l'associer à la capacité à effectuer des inférences *inductives* ou à faire des jugements statistiques. Mais on peut aussi rattacher cette capacité à la rationalité\* logique. Il n'existe pas en fait d'accord unanime chez les logiciens et les philosophes quant à la question de savoir si telle ou telle théorie normative\* incarne ou non la rationalité\*. Notoirement la logique inductive, la théorie de la décision et des probabilités admettent diverses versions distinctes, et on a contesté

qu'elles fussent des *logiques* au sens où l'est, de l'aveu de tous, la logique déductive. Mais pour cette dernière également, la rivalité des logiques intuitionnistes et non classiques (cf. ch.IV) peut menacer sa prétention à incarner *la rationalité\** déductive (cf. Engel 1989 pour des réserves à cet égard). Quoi qu'il en soit, c'est en général par référence à la logique classique que les travaux psychologiques sur la *rationalité\** logique ont été réalisés.

Relativement à chaque type de *rationalité\**, l'étude expérimentale du raisonnement\* met en lumière des déviations importantes par rapport aux modèles normatifs\* respectifs. Dans le domaine de la *rationalité\** instrumentale, les fameux travaux de Kahneman et Tversky (Kahneman, Slovic et Tversky 1982) semblent montrer que les sujets commettent des erreurs élémentaires dans l'estimation de la probabilité d'un événement en estimant plus probable la conjonction de deux événements A et B que A lui-même ("paralogisme de la conjonction"), ignorent la loi des grands nombres, se montrent souvent incapables de reconnaître, à travers deux formulations distinctes d'un même problème, l'identité du problème et de sa solution (phénomène d'"ancrage"), et d'une manière générale, font des erreurs élémentaires en violant les principes bayésiens de la décision (cf. également Davidson 1980, essai 14). Dans le domaine de la *rationalité\** logique, les déviations les plus étudiées ont porté sur la célèbre tâche de sélection\* de cartes de Wason (Wason 1968, Johnson-Laird Legrenzi et Legrenzi, 1972, cf. ch. II *supra*, et pour une revue de ces travaux, Evans 1987, Manktelow et Over 1990), qui repose sur un raisonnement élémentaire à partir d'énoncés conditionnels. Bien que les erreurs constatées varient selon les diverses versions de la tâche de sélection\*, et que leur sens soit sujet à de nombreuses interprétations (cf. *infra*), elles sont systématiques et frappantes. Sans qu'on puisse réellement parler ici d'erreurs de raisonnement, les études portant sur le syllogisme ont montré la difficulté qu'avaient les sujets à manipuler des inférences syllogistiques en dehors des deux premières figures ("effet figural", cf. Johnson-Laird 1983). Dans le domaine des inférences inductives, de nombreux auteurs ont mis en valeur la présence de tendances chez les sujets à confirmer une hypothèse plutôt qu'à chercher à la réfuter (Wason 1960). La question de savoir si c'est une déviation par rapport aux normes de la *rationalité\** logique dépend ici de l'opinion que l'on adopte quant à la

nature du raisonnement inductif (ainsi Popper soutient qu'il n'existe pas de "logique inductive").

Ce qui rend toutes ces erreurs troublantes est le fait qu'en dépit de leurs comportements "déviant" sur ces tâches, les adultes humains restent, dans la majeure partie des cas, parfaitement capables de raisonner déductivement ou inductivement, qu'ils ont bien acquis quelque chose comme une capacité inférentielle, et sont capables de "maximiser leur utilité espérée". La question de savoir si l'on peut, et sur quelles bases, continuer à les appeler "rationnels"\* en l'un ou l'autre des sens envisagés, ou s'il ne faut pas porter un diagnostic plus pessimiste et les tenir comme irrationnels\* a suscité une vaste littérature (Cohen 1981,1986; Manktelow & Over1990; Stich 1990; Evans 1992, Engel 1989).

A la suite de Thagard (1988,123), on peut classer les réactions possibles à cette situation en trois catégories :

- (a) les gens sont idiots. Ils sont tout simplement incapables de suivre les normes logiques appropriées;
- (b) les psychologues sont des idiots. Ils n'ont pas su prendre en compte toutes les variables qui affectent les inférences humaines, et qui, si elles étaient prises en compte, permettraient de montrer que les gens suivent en fait les règles appropriées;
- (c) Les logiciens sont des idiots. Ils évaluent le comportement logique par rapport à des critères normatifs inappropriés.

Dans ce qui suit, j'essaierai de défendre une thèse relativement oecuménique en montrant que, comme le suggère Thagard, tout le monde a sa part d'idiotie, et que rien ne justifie des réponses aussi radicales. Je commencerai par examiner les réponses philosophiques générales qui opposent ce que l'on pourrait appeler un optimisme théorique quant à la rationalité\* humaine à un pessimisme tout aussi théorique (§ 4 et 5), pour ensuite les évaluer (§ 6) et envisager comment les diverses théories psychologiques contemporaines du raisonnement peuvent analyser le problème des erreurs de raisonnement et des déviations, apparentes ou non, dans la rationalité inférentielle des agents ( § 7).

#### **4.LA PRESOMPTION DE RATIONALITE**

On peut distinguer trois sortes d'arguments, de même famille mais distincts, en faveur de la thèse selon laquelle les humains doivent être crédités d'une rationalité\* de principe que les erreurs de raisonnement \*qu'ils commettent ne menacent pas. Chacun de ces arguments consiste à soutenir que toute interprétation\* de leurs croyances et de leurs inférences, déductives ou inductives, doit reposer sur ce que l'on appellera une présomption générale et *a priori* de rationalité sans laquelle il serait tout simplement impossible de leur attribuer des croyances et des états mentaux, et donc d'interpréter leur comportement.

#### ***A. Premier argument: la rationalité a priori des croyances***

Cet argument est très simple. Raisonner\* en général, et raisonner\* déductivement en particulier, suppose que l'on ait des croyances, dont on infère d'autres croyances. Inférer, au sens psychologique du terme, c'est passer d'une croyance\* à une autre croyance\*. Mais l'attribution même de croyances ( et d'autres états intentionnels tels que des désirs ou des souhaits, etc.) à un individu ou à un organisme présuppose que ces croyances soient rationnelles au sens où elles sont cohérentes et closes sous une relation de déduction, et que l'individu suive des schèmes d'action rationnels.\* Par conséquent, l'attribution même d'états intentionnels présuppose que les gens dont on interprète\* les états et les actions soient rationnels\*. On ne peut donc découvrir expérimentalement qu'ils sont irrationnels\* et l'irrationalité\* qu'on est tenté de leur attribuer est soit le fait d'une mauvaise interprétation\* soit due à des facteurs extrinsèques par rapport à l'arrière plan globalement rationnel\* de leurs croyances\* et de leurs actions.

Cet argument a été avancé, sous diverses formes, par Quine (1960), Davidson (1980,1984) et Dennett (1969,1987). Quine le formule dans le contexte d'une conception béhavioriste de la croyance\* comme comportement d'assentiment à une phrase. Mais on peut aussi le formuler dans le contexte d'une conception fonctionnaliste computationnelle des états mentaux du type de celle de Fodor (1987,cf.Engel 1992), d'après laquelle les croyances sont des inscriptions mentales symboliques d'un langage de la pensée. Supposons, avec Quine, que nous soyons en présence du langage inconnu d'une peuplade inconnue, sans avoir de

manuel de traduction à notre disposition ( c'est-à-dire dans une situation de "traduction radicale"\*). Si un indigène prononce en notre présence les sons "*Aivaga*" au moment où passe un éléphant, nous aurons une bonne raison de penser qu'il exprime la croyance\* qu'un éléphant passe dans les environs. Si d'un autre côté notre informateur ne dit pas "*Aivaga*" quand passe un éléphant, mais quand passe un membre éminent du parti socialiste local, nous aurons de bonnes raisons de penser que son énoncé ne porte pas sur le pachyderme en question. Autrement dit, nous supposons que notre observateur a formé ses croyances \*en présence de certains objets parce que ses croyances font partie d'une certaine économie cognitive, par laquelle il perçoit ces objets, lesquels justifient ses croyances\*, etc. En d'autres termes, nous présupposons qu'un certain degré de rationalité \* est présent du simple fait qu'avoir une croyance\* suppose une certaine interaction avec la perception et d'autres états mentaux. De même que nous devons, selon Quine, assigner un degré minimal de rationalité à des croyances isolées, nous devons assigner un degré minimal de rationalité\* à des ensembles de croyances\*, et à leurs liens inférentiels. Supposons que parmi les phrases prononcées par le locuteur indigène, on en relève un certain nombre de la forme "*p blurp q*" et d'autres de la forme "*p glop q*", et supposons que nous relevions les régularités suivantes dans la structure des croyances de ce locuteur :

- (i) "*p*" et "**blurp** *p*" ne sont jamais énoncés en même temps, bien que "*p*" soit quelquefois remplacé par "**blurp** *p*" et vice versa,
- (ii) si "*p glop q*" est déjà énoncé et que si l'indigène énonce "*p*", "*q*" est régulièrement énoncé aussi ,
- (iii) si "*p glop q*" est énoncé, et si "**blurp** *q*" est énoncé, "**blurp** *p*" l'est aussi.

La découverte de ces régularités doit nous induire ordinairement à penser que "*p glop q*" exprime une croyance conditionnelle "si *p* alors *q*" et que "**blurp** *p*" exprime une croyance négative "non *p*". Mais supposons que nous découvriions plutôt les régularités suivantes:

- (a) "*p*" et "**blurp** *p*" sont souvent énoncés simultanément, et il n'y a pas de connexion systématique entre l'apparition de l'un et l'apparition de l'autre;
- (b) si "*p glop q*" est énoncé, etsi "*p*" l'est aussi, alors "*q*" n'est pas énoncé;

(c) si "**p glop q**" est énoncé et si "**q**"l'est aussi, alors "**q**" est énoncé.

Dans ce cas, nous serons pas tentés d'interpréter "**blurp**" comme "ne...pas", ni "**glop**" comme "si...alors", car (a)-(c), à la différence de (i)-(iii) ne correspondent à aucune règle de logique que nous connaissions. Si nous devons traduire systématiquement "**glop**" et "**blurp**" ainsi, et attribuer aux indigènes des croyances \*contradictaires, il vaudrait mieux, soutient Quine, rejeter notre traduction ou la modifier. Le point important n'est pas seulement que nous attribuerions des croyances\* illogiques aux indigènes, mais que nous ne serions même pas capables de reconnaître qu'ils *ont* des croyances\*, si nous ne supposons pas que leurs croyances\* obéissent aux règles de la logique usuelle. En ce sens, comme le dit Quine, "la mentalité prélogique n'est qu'un trait injecté par de mauvais traducteurs". Le principe selon lequel nous devons supposer que les individus dont nous interprétons les croyances sont *a priori* conformes à la logique et rationnelles\* en ce sens est le principe que Quine appelle (à la suite de N.Wilson) le *principe de charité* \*. Ce principe n'est pas propre à Quine ni lié spécifiquement à conception behavioriste de la signification qu'il défend. Il a été défendu aussi par Davidson, qui l'étend jusqu'au point de soutenir que toute interprétation\* des croyances\* doit présupposer que *la plupart* des croyances de l'individu interprété sont vraies. Il a été défendu également par Dennett dans sa théorie des "systèmes intentionnels". Quand nous interprétons le comportement d'un système cognitif, naturel ou artificiel, nous devons nécessairement, selon Dennett, adopter la "posture intentionnelle"\* en supposant que les croyances\* du système sont celles qu'il *devrait* avoir s'il était rationnel, que ses inférences doivent suivre la logique, et que ses désirs sont ceux qu'il devrait avoir s'il était un agent rationnel. Davidson et Dennett insistent sur le fait que la charité \* interprétative\* ou la posture intentionnelle\* ne sont pas seulement des principes méthodologiques utiles pour l'interprétation\*, mais des contraintes *a priori* et nécessaires pour toute attribution de croyances\*. L'argument de la présomption de rationalité n'est pas limité à la rationalité\* logique déductive. Davidson(1980,41, essai 14) a suggéré que des normes\* de rationalité\* pèsent également sur le raisonnement \*inductif ( par exemple "le principe d'information totale": accordez créance aux hypothèses confirmées par toutes les information disponibles),

et sur l'action (comme les principes bayésiens de décision, et le principe parallèle au précédent que Davidson (*ibidem*) appelle "principe de continence" : accomplissez l'action que vous jugez la meilleure sur la base toutes les raisons pertinentes disponibles). Toute évaluation quant à la rationalité \*d'un agent doit donc se faire sur l'arrière plan de conditions normatives d'après lesquelles il ne peut être *que* rationnel, si nous devons le comprendre.

### ***B. Second argument : rationalité et évolution***

Comme le précédent, cet argument a pour but d'établir une présomption nécessaire de rationalité. Mais alors que le précédent parvenait à cette conclusion par le biais des propriétés *conceptuelles* de nos *interprétations* et de nos attributions de croyances, celui-ci s'appuie sur une propriété réelle et empirique des systèmes de croyances humains, le fait que ces systèmes soient le produit de l'évolution\* par sélection naturelle. L'argument consiste à soutenir qu'un système cognitif qui a survécu à la sélection et qui est propre à des organismes qui se sont adaptés (par *fitness*) optimalement à leur environnement doit, de ce fait, être rationnel. L'argument reste le plus souvent, chez les auteurs qui le défendent, peu explicite. Par exemple Dennett (1969) soutient qu'un système intentionnel doit non seulement avoir les propriétés mentionnées plus haut, mais aussi se comporter d'après les besoins biologiques et les désirs qu'il est supposé avoir, étant donné son évolution, et il ajoute que la sélection naturelle "garantit" que la plupart des croyances d'un organisme doivent être vraies, et la plupart de ses stratégies rationnelles (1987, , tr.fr. ). On peut (Stich 1990,63) reconstruire l'argument ainsi :

- (1) l'évolution\*est causée par la sélection naturelle,
  - (2) la sélection naturelle choisit les systèmes cognitifs dont la *fitness* est optimale,
  - (3) de tels systèmes sont rationnels,
  - (4) notre système inférentiel est un produit de l'évolution
- par conséquent (5) les systèmes produits par la sélection naturelle ont une *fitness* optimale et sont en ce sens rationnels\*.

### ***C Troisième argument : l'équilibre réflexif***

Le troisième argument repose sur une idée initialement avancée par Goodman (1955) au sujet de la justification des inférences inductives, mais

on peut l'appliquer aussi aux inférences déductives. L'idée de base est la suivante. Nous justifions une inférence en montrant qu'elle est conforme à des règles générales d'inférence. Mais comment justifions-nous ces règles générales ? Quelle que soit la justification (axiomes évidents, lois de la pensée ou conventions), elle devra faire appel aux règles particulières que nous sommes censés justifier, et sera par conséquent circulaire. Mais selon Goodman, cette circularité n'est pas vicieuse, mais vertueuse : "Une règle est amendée si elle conduit à des inférences que nous ne sommes pas prêts à accepter, et une inférence est rejetée si elle viole une règle que nous ne sommes pas prêts à amender", selon un processus d'ajustement mutuel des règles et des inférences acceptées, et jusqu'à ce qu'un accord ou un équilibre soit atteint entre celles-ci. Appliquant cette idée aux normes morales, Rawls (1970) a proposé de l'appeler "procédure d'équilibre réflexif". L.J. Cohen (1981, 1986) a proposé de l'appliquer aux normes logiques. Nous devons partir des jugements intuitifs et préréflexifs que nous avons sur la validité d'inférences particulières, et chercher, à partir de ces intuitions, à construire une théorie normative\* de la compétence déductive des agents qui essaiera de systématiser ces intuitions de la façon la plus simple possible. Une théorie logique sera justifiée si elle est en équilibre réflexif\*. Appliquée à la psychologie, cette stratégie consistera à dire que le système inférentiel d'un agent humain est justifié s'il est en équilibre réflexif. Pour cela Cohen propose que l'on considère la théorie normative destinée à rendre compte du comportement inférentiel déductif d'un agent comme une théorie de sa *compétence* déductive, par opposition à sa *performance*, au sens chomskyen de cette distinction. Mais puisque la théorie normative en question sera une idéalisation de la compétence déductive et qu'elle devra être en équilibre réflexif, on ne pourra pas, selon Cohen, manquer d'attribuer aux sujets une telle compétence déductive idéale pour rendre compte de leur comportement inférentiel. Quelles que soient les erreurs et les déviations par rapport à cette norme\* que pourront constater les psychologues, tout test expérimental de ces erreurs devra supposer la rationalité\* déductive des sujets. Cela n'implique pas qu'il n'y ait pas, selon Cohen, d'erreurs de raisonnement\*, mais que les erreurs, s'il y en a, affectent la performance et non pas la compétence inférentielle des agents. En ce sens, Cohen soutient que les "erreurs" constatées par les psychologues sont soit dues au choix d'un système normatif\* arbitraire qui

n'est pas réellement en équilibre réflexif\* avec les intuitions des agents, soit dues à des "illusions cognitives" créées par les dispositifs expérimentaux eux-mêmes proposés par les psychologues dans leurs questionnaires.

## 5. RATIONALITE LOGIQUE ET RELATIVISME

Stephen Stich (1990) a critiqué systématiquement les trois arguments en faveur de la présomption nécessaire de rationalité\*. Contre le premier, il fait valoir que le principe de charité\* implique une vision trop idéalisée de la rationalité\* des agents, en présupposant que nous devons toujours les considérer comme rationnels\* et cohérents, alors même que nous constatons couramment qu'ils ne le sont pas. Selon Stich, la théorie de l'interprétation de Quine, Dennett et Davidson implique que nous ne pouvons pas attribuer de croyances\* irrationnelles\* (par exemple contradictoires). Mais c'est au contraire, soutient-il, quelque chose que nous faisons constamment. Nous n'avons aucune difficulté à le faire parce que nous n'interprétons pas les contenus intentionnels d'autrui en lui imposant *a priori* un schème idéal et charitable\* de rationalité, mais parce que nous *simulons* ses états intentionnels en essayant d'imaginer, à partir des nôtres, ce qu'il croirait, désirerait, etc. s'il était à notre place. Cette conception "simulatrice" de l'interprétation (cf. Engel 1992, ch.4) n'implique pas une présomption de rationalité\* des agents.

Contre le second argument, Stich fait valoir que les prémisses (2) et (3) et (4) sont douteuses (sans parler de (1), qui est aussi contesté par certains biologistes). Il n'est pas évident, d'abord, que comme l'asserte (2), l'évolution par sélection naturelle choisisse des systèmes qui sont *fit*. Le phénomène de la pléiotropie, qui fait que par exemple le gène responsable de la fourrure blanche des ours polaires soit aussi responsable de leur albinisme et de leur mauvaise vision, montre que l'évolution ne choisit pas nécessairement la solution optimale. Contre (3), Stich fait remarquer qu'il peut y avoir des circonstances dans lesquelles des croyances *fausses*, ou *irrationnelles\**, ont des effets bénéfiques pour un organisme (1990,61-62). Il n'y a donc pas de lien intrinsèque entre la vérité et la fiabilité, et - en ce sens - la rationalité\* d'un système de croyances\* et d'inférences et sa

survie. Enfin, contre (4), Stich fait valoir que s'il est plausible de supposer que l'évolution est responsable des processus conduisant des organismes à avoir des capacités inférentielles, il ne s'ensuit pas qu'elle soit responsable de la sélection de *tel* système inférentiel plutôt que tel autre, et encore moins des règles *particulières* d'un système inférentiel donné (1990,76-70).

Stich soutient finalement que l'argument de l'équilibre réflexif\* ne peut pas justifier une présomption générale de rationalité\*, parce qu'il n'y a aucune garantie que l'on puisse jamais atteindre un équilibre réflexif\* entre diverses théories normatives\* et des intuitions logiques particulières, ni que cet équilibre réflexif ne puisse pas être atteint à partir de règles d'inférences tout à fait inacceptables (1990,83). Même à supposer, ajoute Stich, que l'on puisse préserver un équilibre réflexif\* pour une classe particulière de gens - des experts- capables de s'entendre sur un ensemble minimal de principes normatifs (les logiciens, par exemple), nous n'avons aucune garantie que cet équilibre idéal puisse être atteint. Selon Stich il n'y a pas d'autre solution que de rejeter l'idée même d'une présomption *a priori* de rationalité\*. L'évaluation de la rationalité\* d'un système inférentiel et des normes qui le régissent est toujours relative à un domaine empirique et à un un objectif particulier. Il n'y a pas d'autre justification de nos normes\* logiques (ou cognitives) que pragmatique, en fonction de l'utilité de ces normes\* pour le domaine en question. Selon ce "relativisme\* pluraliste", il n'y a donc aucune raison de postuler une rationalité de principe qui exclurait la possibilité d'erreurs de raisonnement\*, et nous devons être prêts à admettre, à titre de thèse *empirique*, ce que paraissent établir les psychologues, à savoir que les humains sont, dans un grand nombre de cas, irrationnels\* dans leurs raisonnements\* logiques.

## 6. CHARITE ET RATIONALITE

Que valent ces arguments, qu'il conduisent, comme celui de la présomption de rationalité\*, à ce que l'on peut considérer comme une variante de la thèse optimiste (b) ci-dessus, ou comme celui de Stich, à une variante de la thèse pessimiste (a)?

J'accepterai, en premier lieu, les objections de Stich contre l'argument sélectionniste(B). L'évolution\* par sélection naturelle ne peut suffire à nous faire présupposer la rationalité\* générale des agents.

Considérons alors l'argument (A). Stich a raison de dire que la thèse de la rationalité\* nécessaire des croyances est choquante si elle doit nous conduire à l'idée qu'il est impossible *a priori* d'attribuer à un agent des croyances fausses ou irrationnelles\*. Mais elle n'a pas cette conséquence. Le principe de charité\*, en particulier sous sa forme davidsonienne n'entraîne pas qu'il soit impossible d'attribuer des erreurs, mais que l'erreur *massive* ( sur la plupart des sujets) et l'irrationalité \*totale d'un agent sont hautement improbables, voire impossibles. Il n'implique pas non plus qu'on ne puisse attribuer à un agent des croyances fausses, mais qu'il faut minimiser les erreurs qui sont de prime abord *inexplicables* pour l'interprète (Engel 1992 : 84) Ni Davidson ni Dennett ne soutiennent que les principes de rationalité\* idéale qui gouvernent l'attribution des états intentionnels entraînent qu'un agent dont le comportement résiste systématiquement à toute rationalisation n'ait pas de croyances contradictoires. Au contraire, il en est ainsi, par exemple, pour les croyances conduisant aux états de *self deception* \* (ou duperie de soi) étudiés par Davidson 1985 : si un agent qui se trouve croire que *p* est amené ensuite à croire également que *non p* parce que *p* est pour lui une croyance désagréable, il est victime d'une forme particulière d'irrationalité\* que nous ne sommes nullement autorisés à exclure au nom du principe de charité logique selon lequel l'agent ne *devrait* pas croire à la fois que *p et non p* . Si Davidson devait, en vertu de ce principe, supposer ces croyances cohérentes alors même qu'elles ne le sont pas, il ne pourrait même pas envisager l'existence d'un phénomène comme la *self deception* \*( ou d'autres, comme celui de l'incontinence ou *akrasia* , où l'agent juge qu'il est meilleur de faire *A*, mais ne fait pas *A*, cf. Davidson 1980, ch.2;1982). Il ne s'ensuit nullement que le principe de charité comme condition générale d'attribution des croyances soit infirmé par ce genre de cas, parce que rien, dans la description précédente, ne nous autorise à attribuer à l'agent la croyance conjonctive contradictoire que *p et non p* : l'agent peut croire chacun des énoncés sans croire leur conjonction (Davidson 1982,1991,p.46). Dans ce cas, l'agent a bien, en un sens, des croyances contradictoires, mais cela ne veut pas dire qu'il entretient

*explicitement* une contradiction. En d'autres termes, l'existence d'irrationalités\* particulières, comme celles qui se manifestent dans la *self deception* \*, ne menace pas la validité du principe de charité comme principe général d'attributions de croyances\*.

Stich soutient aussi, comme on l'a vu, que la procédure d'interprétation\* "charitable" ou idéalisante proposée par Quine, Davidson et Dennett est irréaliste, au regard de la procédure d'interprétation "par simulation" qu'il favorise. Mais rien n'indique que les deux procédures soient incompatibles, et qu'il y ait nécessairement une différence entre les croyances\* que l'on attribue à un agent en vertu de ce qu'il *devrait* croire s'il était idéalement rationnel\* et les croyances\* qu'on attribue à cet agent en vertu de ce que l'on *croirait* si l'on se mettait "à sa place" (Dennett 1987,ch.3 ; Engel 1992: 85). Il ne peut pas y avoir une divergence fondamentale entre le principe de charité\* qui prescrit de rationaliser les croyances de l'agent et le principe de simulation qui prescrit de les rendre intelligible. Car l'interprète n'impose pas seulement aux individus interprétés des normes\* transcendantes de rationalité\*, mais aussi ses *propres* normes. Mais alors de deux choses l'une: ou bien l'interprète se tient lui-même comme optimalement rationnel\*, et en ce cas le principe de simulation s'identifie au principe de charité\*, ou bien il se tient lui-même comme rationnel\* jusqu'à un certain degré, et en ce cas le principe de simulation le conduit à traiter l'agent comme *aussi rationnel\* que possible* , ce qui est exactement la stratégie préconisée par la méthode charitable d'interprétation\* de Davidson (Laurier1992).

Ces remarques permettent de faire justice d'une stratégie couramment adoptée par les auteurs qui reprochent aux partisans de la thèse de la présomption de rationalité\* de postuler une rationalité\* trop idéale ou optimale chez les sujets humains peu réaliste d'un point de vue psychologique (Cherniak 1984, Nisbett et Thagard 1983). La rationalité\* humaine, soulignent ces auteurs , est nécessairement *moins* qu'optimale ou idéale, et elle est "limitée" : les agents ont des ressources cognitives finies, ils ne sont pas en mesure d'inférer toutes les conclusions que la logique prescrirait d'inférer et que seuls des êtres omniscients ou peut-être des ordinateurs très puissants seraient en mesure d'inférer. Les modèles de rationalité optimale de la logique déductive usuelle (ou de certaines logiques non classiques, comme la logique épistémique) sont donc

inadéquats. La seule rationalité \* qu'on est en droit d'attribuer *a priori* serait donc, d'après cette conception, une rationalité\* *minimale*, qui ne prescrirait que de tenir les croyances\* des agents comme rationnelles\* *dans un grand nombre de cas mais pas tous*, ou comme *plus ou moins rationnelles\**(Cherniak 1984,ch.1). Le problème posé par cette conception est qu'on ne voit pas bien quels sont exactement les critères de cette rationalité\* minimale, ni quel degré affaibli de logique on doit retenir : combien de principes de rationalité\* parfaite ou idéale faut-il soustraire pour que le seuil approprié de "faisabilité" d'un système inférentiel soit atteint ? Si l'on interprète le principe de charité\* comme on l'a proposé ci-dessus, il n'est pas nécessaire d'envisager des critères de rationalité\* minimale ou limitée : le principe de charité\* est *déjà* un principe de rationalité\* minimale. La méthode d'interprétation\* préconisée par Davidson et Dennett n'implique en rien que l'on ne puisse pas *réviser* à la baisse des attributions de croyances jugées, à des étapes ultérieures de l'interprétation, trop optimistes. Telle est la fonction d'une norme\* de rationalité\* : elle n'est pas destinée à *décrire* l'état de la psychologie d'un individu, mais à poser un *étalon*, une *règle* \*à laquelle la description doit se conformer (Wittgenstein 1956,Bouveresse 1987).En ce sens il est absurde de supposer que les principes normatifs\* de rationalité\* devraient décrire, comme des cartes décrivent un territoire, le réseau des croyances d'un individu.Les normes\* sont des règles\* pour la description, elles n'ont pas elles-mêmes le statut de descriptions.

Considérons, enfin, l'argument (C) de l'équilibre réflexif\*. Il me paraît fondamentalement correct, en dépit des critiques de Stich (Engel 1989,ch.XIII). Selon cet argument, tout système qui passe le test de l'équilibre réflexif \**est rationnel\**. Cela désarme des objections comme celle de Stich, pour qui des schèmes de raisonnement\* irrationnels\* pourraient, s'ils étaient approuvés à la fois par les intuitions des sujets et conformes aux règles normatives\* qu'ils se donnent. Mais tout dépend de la portée des principes et des jugements mis en jeu dans l'équilibre réflexif\*. Pouvons-nous nous fier à une majorité de jugements de profanes, et devons-nous ignorer les principes des logiciens et des théoriciens des probabilités qui nous montreraient que les schèmes établis sont contradictoires ? Si nous élargissons la base des jugements et des principes, nous obtiendrons un équilibre réflexif\* différent. A nouveau, cet

équilibre pourra valider des règles que la logique usuelle et les experts sanctionneront comme incorrectes. Mais d'une manière ou d'une autre, nous utilisons pour justifier nos règles une procédure d'évaluation quelconque, conduisant à un équilibre entre nos jugements intuitifs et nos règles. Le fait même d'avoir une procédure de ce genre *est* ce que nous appelons une justification de nos règles d'inférence, et si nous en avons une nous sommes rationnels. Comme le remarque Thagard (1988:121), l'équilibre réflexif\* peut être "étroit" et n'aller que des intuitions aux théories normatives particulières, ou "large" et inclure aussi l'évaluation de théories d'arrière plan empiriques (par exemple des théories psychologiques sur les capacités inférentielles des sujets). Tous ces éléments doivent entrer en ligne de compte dans l'équilibre atteint. Peut-on, comme Stich, supposer que diverses communautés, ou divers domaines d'intérêt puissent conduire à des conflits entre l'équilibre réflexif atteint par un groupe et celui atteint par un autre ? C'est une possibilité qui paraît plus théorique que réelle. Car un conflit profond entre des équilibres réflexifs distincts signifierait que l'on a affaire à des schèmes de pensée non traductibles ou incommensurables, comme le conflit allégué entre une mentalité "logique" et une mentalité "prélogique". Mais si le principe de charité\* interprétative est correct, il y a peu de chances pour que l'on puisse assister à une véritable divergence, comme l'a soutenu notamment Davidson (1984,ch.13). On répondra qu'un tel conflit semble exister, avec le cas de la logique intuitionniste se posant en rivale de la logique classique. Mais peut-on dire que ce conflit recouvre deux schèmes de pensée incommensurables ? C'est douteux, car nous pouvons *comparer* les principes de l'une et l'autre logique, et on dispose de critères logiques de comparaison (comme la non contradiction, et par conséquent d'une base commune d'évaluation nous permettant de savoir *quelle* conception de la déduction est justifiée (cf. Engel 1989,ch.XII). En d'autres termes les conflits entre logique classique et logique non classique sont des conflits internes à la logique, qui présupposent qu'on ait atteint un équilibre réfléchi.

On doit donc, à mon sens, admettre la thèse de la présomption de rationalité, en tous cas sous sa forme (A) et (C). S'ensuit-il que cette thèse règle notre problème initial, celui de l'explication des erreurs inférentielles constatées par les psychologues et qu'elle suffise à répondre au problème

de la rationalité inférentielle? C'est douteux. Rappelons qu'une des prémisses de l'argument (*A*) est que le raisonnement\* est une transition entre des croyances\*. Dans la mesure où prêter à un sujet un certain raisonnement c'est lui attribuer certaines croyances\*, l'argument conduit à dire qu'il ne peut être *que* rationnel. Mais comme le montre le cas mentionné ci-dessus de raisonnements\* ou de conduites irrationnelles\* comme la *self deception*, le fait que les agents soient *en droit* rationnels n'entraîne pas qu'ils ne puissent pas occasionnellement obéir à des schèmes qui ne le sont pas. La thèse de la présomption de rationalité\* des croyances est trop générale pour nous permettre d'expliquer les formes particulières que peuvent prendre aussi bien la rationalité que l'irrationalité de nos schèmes inférentiels. Ici seule l'étude psychologique empirique peut nous renseigner, et l'argument conceptuel développé jusqu'ici n'affecte pas les recherches empiriques sur l'irrationalité.

On peut développer également un autre argument en vue de soutenir que la thèse de la présomption de rationalité n'est pas pertinente pour les recherches empiriques sur le raisonnement (Jacob 1991). Les théories psychologiques cognitives contemporaines du raisonnement ne traitent pas, en général le raisonnement\* logique comme une simple transition entre des croyances\*, mais comme une manipulation de processus et de représentations cognitifs *infradoxastiques* ou *subdoxastiques* (cf. Stich 1978). Ces processus et représentations n'ont pas, de prime abord, le caractère de contenus intentionnels ou propositionnels entretenus consciemment ou directement accessibles à la conscience ou à la verbalisation des sujets. En ce sens ils sont "subpersonnels" ou "subrationnels" (Dennett 1987, ch.2). Peut-être faut-il, pour les distinguer des croyances proprement dites, les appeler des "proto-croyances" (par exemple Jacob 1991 distingue les "proto-croyances" que forme spontanément un sujet quand il comprend une phrase dans un certain contexte, des croyances qu'il peut entretenir au sujet de cette phrase sans en connaître le contexte). Mais en ce cas la question de leur rationalité, si elle se pose, ne peut pas se poser pour eux de la même manière que pour les croyances ordinaires. L'argument de la rationalité *a priori* des croyances, selon cette analyse, ne peut donc pas s'appliquer à ces représentations et processus. Il n'a, soutient Pierre Jacob, "*aucune incidence sur les recherches expérimentales sur le raisonnement*"

(1991,209). Cette conclusion très forte dépend de la possibilité de distinguer nettement les croyances conscientes, réfléchies, ou explicites sur lesquelles porte l'argument (*A*) des états subdoxastiques ou des proto-croyances qui interviennent dans le raisonnement effectif des sujets. Or cette distinction dépend elle-même du type de théorie psychologique du raisonnement que l'on adopte. C'est donc vers les différents modèles qu'ont proposé les psychologues qu'il faut finalement se tourner pour répondre à cette question.

## 7.PSYCHOLOGIE DU RAISONNEMENT ET RATIONALITE

On ne peut ici examiner toutes les théories en présence (cf. le ch. II de ce volume). Je considérerai principalement les modèles du raisonnement déductif, en prenant, comme c'est souvent le cas dans ces discussions, comme référence les résultats de la tâche de sélection\* de Wason. Les principales théories contemporaines sont la théorie de la logique mentale (Braine 1990), celle des modèles mentaux (Johnson-Laird 1983, Johnson Laird & Byrne 1991), celle des schémas pragmatiques (Cheng & Holyoak 1985) celle des contrats sociaux (Cosmidès 1989), et celle des biais et heuristiques (Evans 1989).

La théorie psychologique du raisonnement qui s'accorde le mieux avec une distinction entre croyances ordinaires et états subdoxastiques est sans doute celle de la logique mentale\* (Henle 1962, Braine 1990). Selon cette conception, les gens raisonnent en appliquant les règles d'une logique mentale\* interne, qui sont principalement des règles de déduction naturelle comparables à celles des systèmes de logique du même nom. Ils appliquent ces règles de manière automatique sans former au sujet des inférences qu'ils effectuent des croyances conscientes. Dans ce cas, la présomption de rationalité, qui s'applique aux croyances conscientes propositionnelles, ne peut pas s'appliquer aux processus d'inférence gouvernés par les règles de la logique mentale. Comment, selon cette conception, expliquer les erreurs inférentielles? Car si les individus *ont* une logique interne, comment peuvent-ils se tromper ? Pour répondre à cette question, les défenseurs de

la logique mentale\* distinguent en général les *processus* (subdoxastiques) par lesquels les sujets manipulent des représentations internes pour effectuer des inférences, des manières dont ces représentations sont *interprétées* par les sujets (Henle 1962). Ils insistent aussi sur le caractère sémantique et pragmatique de ces processus d'interprétation, par opposition au caractère syntaxique des règles de la logique mentale (Politzer 1992). Ce sont ces interprétations qui sont responsables des erreurs. Selon cette hypothèse, par conséquent, les gens raisonnent correctement, mais se représentent faussement les données du problème. Cette conception revient à attribuer une rationalité aux agents, mais ce n'est pas celle visée par l'argument (A) : c'est celle du système inférentiel dont les agents disposent, qu'on peut juger plus ou moins économique, plus ou moins efficace cognitivement, mais auquel des jugements normatifs, quant à son caractère correct ou justifié ne peuvent s'appliquer.

Mais l'hypothèse de la logique mentale est loin d'être la seule possible. Selon Johnson-Laird (1983) il peut y avoir du raisonnement, et du raisonnement correct, sans logique mentale ni règles d'inférence. Selon la théorie des modèles mentaux\*, les individus construisent des modèles ou représentations sémantiques des prémisses et de la conclusion d'une inférence déductive, et les évaluent sémantiquement de manière à trouver les contre-exemples possibles à ces inférences. Les erreurs de raisonnement interviennent quand la complexité de la tâche de représentations de modèles est trop grande et requiert des ressources mémorielles importantes ("effet figural" des syllogismes, cf. ci-dessus), ou, dans le cas des raisonnements conditionnels, quand ils forment des modèles appropriés des antécédents (positifs ou négatifs) des énoncés conditionnels (Johnson-Laird and Byrne 1991). La construction de modèles mentaux par les sujets est-elle une activité consciente et réfléchie, qui donne lieu à des croyances authentiques sur les propriétés des énoncés manipulés, ou bien est-ce une activité irréfléchie, plus ou moins automatique, qui repose sur des processus subdoxastiques ? La réponse est mitigée : d'un côté une bonne partie des processus de formation des modèles mentaux sont implicites (par exemple dans la mémoire), mais de l'autre Johnson-Laird insiste sur le fait qu'il y a une *évaluation* sémantique des prémisses et de la conclusion par le sujet. Johnson-Laird et Byrne (1991) soutiennent également que les croyances et les

connaissances (au sens usuel) des sujets influent sur le processus de déduction, alors que les règles formelles d'une logique mentale sont "aveugles" aux croyances que les sujets entretiennent sur les contenus des prémisses. Dans une telle conception, il *est* pertinent d'attribuer des croyances "rationnelles" *a priori* aux agents. Et de fait Johnson-Laird et Byrne (1991, 1993), quand ils considèrent le problème de la rationalité, développent l'idée qu'il existe un "noyau" de rationalité commun à tout agent humain qui raisonne, à savoir le principe de validité sémantique, selon lequel un argument est valide si les prémisses étant vraies, la conclusion ne peut être fausse. Cette capacité "universelle" des agents humains peut être considérée comme une présomption nécessaire attribuable à tout organisme capable de raisonner sémantiquement, par construction de modèles (Engel 1993).

De nombreux travaux ont mis en valeur l'influence des contenus particuliers des tâches proposées par les expérimentateurs, c'est à dire des croyances\* des agents au sujet de ces contenus, sur les performances inférentielles. Ainsi la tâche de sélection\* paraît facilitée, dans certaines circonstances, par la présence de contenus réalistes (Johnson-Laird 1983). Evans (1992) fait remarquer que parmi deux syllogismes formellement équivalents, l'un est jugé (correctement) non valide quand le contenu est difficile à croire, alors que l'autre est (incorrectement) jugé valide quand le contenu est plus crédible. Différents chercheurs ont suggéré que les performances inférentielles sont déterminées par l'emploi de règles dotées de contenus spécifiques, et en particulier que certains types de contextes thématiques évoquent des "schèmes pragmatiques"\* (ou représentations stéréotypiques de situations) qui permettent aux sujets de raisonner ( Cheng & Holyoak 1985). Ces schémas correspondent à des règles\* de production définies en termes de classes d'objectifs ou d'obligations particuliers. Cosmides (1989) a soutenu que les schèmes en question étaient des "contrats sociaux"\* innés, produit de l'évolution\* de l'espèce à travers ses échanges sociaux. Cette dernière hypothèse pourrait être rattachée à l'argument évolutionniste\* (B) ci-dessus en faveur de la rationalité nécessaire des agents. Mais on a vu tout ce que cet argument avait de douteux. La question importante reste donc la même que précédemment : peut-on, dans l'une ou l'autre de ces hypothèses, distinguer nettement des processus subdoxastiques des croyances proprement dites?

Certes des structures innées produites par l'évolution ne sont pas des croyances au sens ordinaire, et les schémas pragmatiques sont des règles que, selon cette hypothèse, les sujets utilisent implicitement et de façon spontanée et non pas explicitement. Mais ce qu'ont en commun la théorie des schémas pragmatiques et celle des contrats sociaux n'est pas seulement l'hypothèse de règles à contenu spécifique, mais l'idée que le raisonnement déductif humain est étroitement lié aux croyances que les sujets entretiennent sur le contexte. Pour qu'un module inférentiel activant par exemple un schéma pragmatique puisse se mettre en marche, il faut que le sujet puisse extraire de la situation qu'on lui propose diverses données à partir de ses connaissances générales.

Considérons enfin la théorie des "biais"\* proposée en particulier par Evans (1989) pour le raisonnement déductif, et pour le raisonnement inductif par Kahneman et Tversky (1982). Ces derniers ont soutenu que les diverses erreurs inférentielles constatées étaient dues à des biais\* spécifiques en faveur de telle ou telle caractéristique saillante du problème présenté par l'expérimentateur ( par exemple la similarité des items dans une évaluation de probabilité, ou la correspondance ("matching") des items dans la tâche de sélection\*). Pour raisonner les sujets, selon cette hypothèse, utilisent des "heuristiques\* représentationnelles" qui guident leurs réponses erronées ou correctes. Ici encore, ces biais\* sont largement implicites, non verbaux et pré-attentifs. Mais l'usage de ces biais n'empêche nullement, selon Evans, les sujets d'employer par ailleurs des processus "analytiques", conscients et réfléchis de raisonnement à partir des traits d'information sélectionnés antérieurement par divers biais. De même dans le cas du raisonnement inductif, certains auteurs suggèrent la co-existence, chez les sujets, de schèmes de raisonnement bayésiens et d'heuristiques\* non conformes à ces schèmes (Osherson 1990).

Il n'entre pas dans mon propos de comparer les mérites de ces diverses théories (cf. Politzer et Nguyen Xuan 1992). Mais il ressort de ce rapide passage en revue que si la plupart des modèles psychologiques proposés postulent l'existence de règles et de processus inférentiels subdoxastiques, elles n'excluent pas non plus l'usage, dans le raisonnement, de croyances réfléchies et conscientes, ni l'évaluation sémantique de prémisses et de conclusion par application de règles méta-inférentielles (portant *sur* les propriétés des inférences à accomplir). Il s'ensuit que l'argument indiqué à

la fin de la section précédente, selon lequel la thèse de la présomption de rationalité ne peut pas s'appliquer aux processus subdoxastiques et subrationnels du raisonnement logique mis en valeur par les psychologues, et n'est pas pertinent pour l'évaluation de la rationalité des sujets, n'est que partiellement justifié et a une portée limitée. Il ne vaut, en toute rigueur, que dans l'hypothèse où le raisonnement logique s'effectue par l'intermédiaire de l'activation de règles\* inférentielles largement inconscientes, automatiques, et modulaires. Mais rien n'interdit de dire que la présomption de rationalité s'applique pour les croyances conscientes que les sujets utilisent quand ils font des inférences, soit par les processus que Evans appelle analytiques, soit par les évaluations sémantiques de modèles mentaux dont parle Johnson-Laird. En d'autres termes, même s'il est vrai que de nombreux cas de raisonnements courants relèvent de l'usage de règles spécifiques apprises relativement à un domaine spécifique dont le fonctionnement ne suppose pas l'existence d'une compétence rationnelle générale, il ne s'ensuit pas que tout raisonnement s'effectue uniquement au moyen de telles règles spécifiques et soit indépendant d'une telle compétence générale. On ne peut pas exclure non plus que les erreurs inférentielles constatées reposent dans certains cas sur l'emploi d'une théorie normative inappropriée (ce qui tendrait à renforcer la stratégie (C) ci-dessus). Par exemple Gigerenzer (1991) a suggéré que nombre des erreurs attribuées dans les évaluations de probabilité peuvent disparaître quand on suppose qu'ils usent de règles fréquentistes de probabilité, et non pas de règles\* subjectivistes ou bayésiennes\*, et qu'ils sont en ce sens, beaucoup plus rationnels\* que ne le laissent supposer les résultats courants de la recherche sur les biais\*.

Un certain nombre des difficultés qu'on rencontre quand on se demande si les agents raisonnent selon des règles rationnelles\* ou non procèdent en fait d'une difficulté fondamentale portant sur la notion même de *règle\**. Peut-on dire, comme le soutiennent notamment les partisans de la thèse de logique mentale\*, que les agents utilisent, ou suivent, des règles\* d'inférence non conscientes, ou bien faut-il limiter l'emploi du mot "règle\*" à un épisode conscient, consistant à appliquer tel ou tel schème de manière réfléchi? Wittgenstein (1956) et ses disciples (Kripke 1981) soutiennent que la question de la normativité\* des règles\* (et de la logique) n'a de sens que pour les règles\* du second type, et pas pour les

règles\* du premier type. Il ne s'ensuit pas que l'emploi du mot "règle\*" au premier sens soit illégitime, mais que la question posée est distincte.

La question de la rationalité et de la normativité\* de la logique n'est donc pas une question à laquelle on puisse donner une réponse unilatérale, parce que, comme on l'a vu, elle recouvre différentes questions. L'argument de la présomption de rationalité\* attribuable à tout sujet capable d'avoir des croyances est fondamentalement correct, et en ce sens il paraît difficile d'envisager que les recherches empiriques sur le raisonnement puissent jamais démontrer l'irrationalité\* des agents. Cela, comme on l'a vu, n'a rien de surprenant, puisqu'une norme\* de rationalité\* n'a pas, en elle-même, de pouvoir descriptif sur la psychologie des agents, mais essentiellement une force prescriptive. La question de savoir si des agents suivent des normes\* générales de rationalité\* est donc distincte des hypothèses particulières que l'on peut faire sur les règles inférentielles qu'ils suivent psychologiquement (Engel 1993a). L'argument de la présomption de rationalité\* n'établit pas que les humains soient en tous points rationnels\* dans leurs raisonnements\*, et ne permet en rien de déterminer quels types de processus ils mettent en oeuvre dans leur pratique courante du raisonnement\*. La conclusion qu'il faut en tirer, semble-t-il, est qu'il n'y a pas plus de raisons de se réjouir que les gens soient toujours rationnels\* que de raisons de se lamenter qu'ils ne le soient pas.

### Glossaire-index (termes marqués"\*)

biais ; charité (principe de); contrats sociaux; croyance; équilibre réflexif; états subdoxastiques; évolution, évolutionnisme ; heuristique; interprétation ; irrationalité, irrationnel; logique mentale; modèles mentaux; norme, normativité; posture intentionnelle; psychologisme, anti-psychologisme; raisonnement; rationalité, rationnel; règle; schéma pragmatique; *self deception*; tâche de sélection; traduction radicale

### Références

P.Engel, *Logique, raisonnement et normes de rationalité*, à paraître in O. Houdé et D. Miéville, *Pensée logico-mathématique, nouveaux objets interdisciplinaires*, Paris, PUF 1993

- Bouveresse, J. (1987), *La force de la règle, Wittgenstein et l'invention de la nécessité*, Paris, Minuit
- Braine, M. (1990) "The "Natural Logic" Approach to Reasoning", in W.. Overton, ed., *Reasoning, necessity, and logic : Developmental Perspectives*, Hillsdale, N.J., L. Erlbaum
- Cheng, P. et Holyoak K.J (1985) "Pragmatic Reasoning Schemas" *Cognitive Psychology*, 17, 391-416
- Cherniak, C. (1986) *Minimal Rationality*, Cambridge, Mass, MIT Press
- Cohen, L.J. (1981) "Can Human Irrationality be Experimentally Demonstrated?", *Behavioral and Brain Sciences*, 4, 317-70
- Cohen, L.J. (1986) *The Dialogue of Reason*, Oxford, Clarendon Press
- Cosmides, L. (1989) "The Logic of Social Exchange : Has Natural Selection Shaped How Humans Reason ? Studies on the Wason Selection Task", *Cognition*, 31, 187-328
- Davidson, D. (1980), *Essays on Actions and Events*, Oxford, Oxford University Press, tr. fr. P. Engel, *Actions et Evenements*, Paris, P.U.F. 1993
- Davidson, D. (1982) "Paradoxes of Irrationality", in J. Hopkins et R. Wohlheim, eds, *Philosophical Essays on Freud*, Cambridge, Cambridge University Press,  
tr. fr. in Davidson 1991
- Davidson, D. (1982a) "Rational Animals", in Le Pore ed. 1985, tr. fr. in Davidson 1991
- Davidson, D. (1984) *Inquiries into Meaning and Truth*, Oxford, Oxford University Press, tr. fr. P. Engel, à paraître, ed. J. Chambon
- Davidson D. (1985) "Deception and Division" in Le Pore ed. 1985, tr. fr. in Davidson 1991
- Davidson, D. (1991) *Paradoxes de l'irrationalité*, tr. fr. P. Engel, Combas, l'Eclat
- Dennett, D. (1969) "Intentional Systems", *The Journal of Philosophy*, repris dans D. Dennett 1978, tr. fr. J. Kahlfa, in *Philosophie* 1, 1984
- Dennett, D. (1978), *Brainstorms*, Bradord Books, MIT, Cambridge Mass
- Dennett, D. (1987), *The Intentional Stance*, Cambridge Mass, MIT Press, tr. fr. P. Engel, *La stratégie de l'interprète*, Paris, Gallimard, 1990
- Elster, J. (1982), "Rationality", in G. Floistad, ed. *contemporary Philosophy*, vol.2, La Haye, Nijhoff, 111-131
- Engel, P. (1989) *La norme du vrai, philosophie de la logique*, Paris, Gallimard, ed. anglaise corrigée et révisée par P. Engel et M. Kochan, *The Norm of Truth, an Introduction to the Philosophy of Logic*, Hemel Hempstead, Harvester Wheatsheaf, 1991
- (1991), "Raisonnement déductif et rationalité", ms, Ecole d'Eté de l'ARC (Bonas), Cahiers et Rapports du CREA.
- (1992), *Etats d'esprit, questions de philosophie de l'esprit*, Aix en Provence, Alinéa.

P.Engel, *Logique, raisonnement et normes de rationalité*, à paraître in O. Houdé et D. Miéville, *Pensée logico-mathématique, nouveaux objets interdisciplinaires*, Paris, PUF 1993

- (1993 )"Mental Model Theory and Rationality", à paraître,*Behavioral and Brain Sciences*.
- (1993a) "Three Forms of Normativity, a reply to Richard Nisbett ", ms, tr.fr. à paraître in *Lire Davidson*, Combas, L'Eclat
- Evans, J.st.B. (1989)*Biases in Human Reasoning, Causes and Consequences*, Hove and London, Erlbaum
- (1987 ) "Reasoning", in H. Beloff and A.Colman,*Psychology Survey*, The British Psychological Society
- (1992) "Bias and Rationality", in Manktelow and Over 1992
- Fodor, J. (1987) *Psychosemantics*, Cambridge Mass, MIT Press
- Frege, G. 1893 *Grundgetetze der Arithmetik*, vol. 1, Breslau, re. Olms
- Gigerenzer, G.( 1991) "How to Make Cognitive Illusions Disappear : Beyond Heuristics and Biases", in W. Stroeber & M Hewstone, eds. *European Review of Social Psychology*, vol.2, New York, Wiley.
- Goodman , N. (1955)*Fact, Fiction and Forecast*, Bobbs Merrill, New York, tr. fr. P.Jacob et alii, *Faits, fictions et prédictions*, Paris, Minuit 19
- Henle, M.(1962) "On the Relation between Logic and Thinking", *Psychological Review*, 69, 366-378.
- Husserl, E. (1901) *Logische Untersuchungen*, Halle, Max Niemeyer, tr.fr.A. Kelkel et R. Schérer, *Recherches logiques*, Paris, PUF 1962
- Jacob, P. (1991),"Interprétation,inférence et rationalité",*Dialectica*, 11,193-217
- Jeffrey, R.(1965 )(2ème ed.1980)*The Logic of Decision*, Chicago, Chicago University Press
- Johnson-Laird,P. Legrenzi, P.et Legrenzi ,M.(1972) "Reasoning and a Sense of Reality", *British Journal of Psychology*, 63, 395-400
- Johnson-Laird, P.(1983 ) *Mental Models*, Cambridge University Press
- Johnson-Laird, P. and Byrne, R.(1990) *Deduction*, Hove and London, Erlbaum
- Johnson-Laird, P. and Byrne, R.( 1993) "Précis" of *Deduction, Behavioral and Brain Sciences*, à paraître, 1993
- Kahneman, D.Slovic, P. and Tversky, A. (1982) eds, *Judgment under Uncertainty, Heuristics and Biases*, Cambridge, Cambridge University Press
- Kripke, S.(1981) *Wittgenstein on Rules and Private Language*, Blackwell, Oxford
- Laurier, D.(1992) "Rationality and Intentionality: a Defense of Optimization in Theories of Interpretation", *Cahiers du Département de Philosophie de l'Université de Montréal*, n° 9207
- Le Pore, E. (1985) ed.,*Actions and Events : Perspectives on the Philosophy of Donald Davidson*, Oxford, Blackwell
- Manktelow, K.I., et Over, D. (1990) *Inference and Understanding*, Routledge, London
- Mc Namara, J. (1986)*A Border Dispute,the Place of Logic in Psychology*, Cambridge Mass, MIT Press
- Nisbett, R. et Thagard, P.(1983) "Rationality and Charity", *Philosophy of Science*, 50,

P.Engel, *Logique, raisonnement et normes de rationalité*, à paraître in O. Houdé et D. Miéville, *Pensée logico-mathématique, nouveaux objets interdisciplinaires*, Paris, PUF 1993

- Osherson, D. et alii ( 1990) *An Invitation to Cognitive Science*, 3 vol. Cambridge, Mass,  
MIT Press
- Osherson, D.(1990 )" Judgment", in Osherson et alii, vol.3, *Thinking*, 55-87
- Politzer, G. (1992) "Logique mentale et raisonnement naturel" in D.Andler et alii, eds, *Epistemologie et cognition*, Bruxelles, Mardaga, 79-86
- Politzer, G. et Nguyen Xuan, A. (1992) " Reasoning about Conditional Promises and Warnings : Darwinian Algorithms, Mental Models, Relevance Judgments or Pragmatic Schemas ?", *Quarterly Journal of Experimental Psychology*, 44 A(3), 4101-421
- Quine, W.V.O. 1960, *Word and Object*, Cambridge Mass, MIT Press
- Rawls, J (1970), *A theory of Justice*, Harvard, Harvard University Press, tr.fr. C.Audard, *Théorie de la justice* ,Paris, Seuil, 1986
- Stich, S.(1978) "Beliefs and Subdoxastic States", *Philosophy of Science*, 45, 499-518
- Stich, S. (1990) *The Fragmentation of Reason*, Cambridge Mass, MIT Press
- Thagard, P. (1990) *Computational Philosophy of Science*, Cambridge, Mass, MIT Press
- Wason, P. (1960) "On the Failure to Eliminate Hypotheses in a Conceptual Task", *Quarterly Journal of Experimental Psychology*, 12, 129-140
- Wason, P.(1968) "Reasoning about a Rule", *Quarterly Journal of Experimental Psychology*, 20, 273-81
- Wittgenstein,L.(1956) *Bemerkungen über die Grundlagen der Mathematik*, Oxford, Blackwell, , tr.fr. M. Lescourret, *Remarques sur les fondements des mathématiques* , Paris,Gallimard